СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

УДК 519.7 Обзорная статья
DOI: 10.35330/1991-6639-2023-6-116-152-159
EDN: LBDSYZ

Интеллектуальные методы кластеризации данных

Р. А. Жилов

Институт прикладной математики и автоматизации –
филиал Кабардино-Балкарского научного центра Российской академии наук
360000, Россия, г. Нальчик, ул. Шортанова, 89 А

Аннотация. В данной работе рассматриваются интеллектуальные методы кластеризации данных. В последние годы наблюдается увеличение количества данных, которые подлежат анализу в различных областях. В связи с этим возрастает потребность в более эффективных методах кластеризации данных. Методы кластеризации данных можно разделить на две основные категории: иерархические и неиерархические. Иерархические методы кластеризации строят дерево кластеров, начиная с каждого объекта в отдельном кластере, а затем объединяют близкие кластеры, пока не останется один кластер, содержащий все объекты. Неиерархические методы кластеризации определяют число кластеров заранее и группируют объекты в соответствии с их сходством и различиями. Методы кластеризации данных – это одна из важнейших областей машинного обучения, которая позволяет группировать данные в соответствии с их признаками и характеристиками. Кластеризация данных является одним из основных методов анализа данных и находит широкое применение во многих областях, включая биологию, медицину, экономику, социологию и другие.

Ключевые слова: кластеризация данных, метод k-средних, метод DBSCAN, методы кластеризации на основе плотности, метод SOM

Поступила 24.10.2023, одобрена после рецензирования 02.11.2023, принята к публикации 09.11.2023

Для цитирования. Жилов Р. А. Интеллектуальные методы кластеризации данных // Известия Кабардино-Балкарского научного центра РАН. 2023. № 6(116). С. 152–159. DOI: 10.35330/1991-6639-2023-6-116-152-159

MSC: 68T09 Review article

Intelligent data clustering methods

R.A. Zhilov

Institute of Applied Mathematics and Automation –
branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences
360000, Russia, Nalchik, 89 A Shortanov street

Abstract. The paper considers intelligent methods of data clustering. In recent years there has been an increase in the amount of data to be analyzed in various fields. As a result, there is a growing need for more efficient data clustering methods. Data clustering methods can be divided into two main categories: hierarchical and non-hierarchical. Hierarchical clustering methods build a tree of clusters, starting with each feature in a separate cluster and then merging close clusters until there is one cluster containing all the features. Non-hierarchical clustering methods determine the number of clusters in advance and group objects according to their similarities and differences. Data clustering methods is one of the most important areas of machine learning, which allows you to group data according to their features and characteristics.

Data clustering is one of the main methods of data analysis and is widely used in many fields, including biology, medicine, economics, sociology, and others.

Keywords: data clustering, k-means method, DBSCAN method, density-based clustering methods, SOM method

Submitted 24.10.2023, approved after reviewing 02.11.2023, accepted for publication 09.11.2023

For citation. Zhilov R.A. Intelligent data clustering methods. *News of the Kabardino-Balkarian Scientific Center of RAS.* 2023. No. 6(116). Pp. 152–159. DOI: 10.35330/1991-6639-2023-6-116-152-159

Введение

Одним из наиболее популярных методов кластеризации данных является метод k-средних. Он относится к неиерархическим методам и представляет собой алгоритм, который разделяет данные на кластеры, основываясь на сходстве объектов друг с другом. Метод k-средних используется для анализа различных типов данных, включая числовые и категориальные.

Еще одним популярным методом кластеризации данных является метод DBSCAN. Он также относится к неиерархическим методам и используется для кластеризации данных, основываясь на плотности объектов в пространстве признаков. Метод DBSCAN может обнаруживать кластеры любой формы и размера, а также распознавать шумовые точки и выбросы.

Кроме того, существуют и другие методы кластеризации данных, такие как методы иерархической кластеризации (например, метод Ward или метод евклидовой агломерации), методы на основе иерархических байесовских моделей (например, Dirichlet Process Gaussian Mixture Models), методы кластеризации на основе плотности (например, Mean Shift) и многие другие [1, 2].

Однако независимо от выбранного метода кластеризации необходимо учитывать ограничения и особенности данных, которые подвергаются анализу. Важно также выбирать правильные параметры и метрики для оценки качества кластеризации, чтобы получить оптимальный результат.

Таким образом, методы кластеризации данных имеют большую важность и ценность в области анализа данных, позволяя находить закономерности и структуры в больших объемах данных. Развитие новых методов кластеризации данных является активной областью исследований, что позволяет получать более точные и эффективные результаты.

Метод k-средних

Метод k-средних (k-means) является одним из самых распространенных методов кластеризации данных. Он позволяет разбить набор данных на заранее определенное число кластеров (групп), где каждый кластер содержит объекты, близкие по своим характеристикам [3].

Алгоритм k-средних работает следующим образом:

Начальное состояние. Определяется число кластеров k, а затем выбираются случайным образом k точек из набора данных. Эти точки становятся начальными центроидами (центрами) кластеров.

Присвоение кластера. Каждый объект данных присваивается к кластеру, чей центройд находится ближе всего к данному объекту.

Обновление центройдов. Для каждого кластера вычисляется новый центроид, который представляет собой среднее значение всех объектов в данном кластере.

Повторение шагов 2 и 3. Шаги 2 и 3 повторяются до тех пор, пока центройды не перестанут изменяться или пока не будет достигнуто максимальное число итераций.

Результат. Полученные кластеры могут быть использованы для анализа данных, классификации объектов и других приложений.

Метод k-средних имеет несколько преимуществ, таких как простота реализации и высокая скорость работы. Однако он также имеет недостатки, например, чувствительность к начальному выбору центройдов и неспособность работать с нелинейно разделимыми данными.

В целом метод является эффективным и широко используемым при кластеризации данных, он может быть применен в различных областях, включая машинное обучение, анализ данных, геоинформатику, биоинформатику и другие.

Метод k-средних широко применяется в различных областях, где требуется кластеризация данных. Например, может использоваться для анализа социальных сетей, обработки и анализа медицинских данных, классификации текстовых документов и других задач.

Одним из примеров использования метода k-средних является задача сегментации клиентов. В этой задаче требуется разбить клиентов на группы схожих по своим характеристикам (например, возраст, доход, предпочтения и т.д.). Это может быть полезно для разработки более эффективных маркетинговых стратегий и улучшения обслуживания клиентов.

Также метод k-средних может быть использован в области обработки изображений, например, для сегментации изображений на отдельные части или объекты. В этом случае каждый пиксель изображения может быть рассмотрен как отдельный объект, а кластеризация данных позволит определить сходство пикселей и разделить изображение на соответствующие кластеры.

Метод также может использоваться в задачах машинного обучения, например, для кластеризации признаковых описаний объектов в задачах классификации и кластеризации. В этом случае метод k-средних может быть использован для построения скрытых представлений данных, которые могут применяться в дальнейшем для обучения модели.

Как правило, метод k-средних работает лучше всего в случаях, когда данные являются числовыми и имеют простую структуру. Метод k-средних может быть менее эффективным в случае, когда данные имеют сложную структуру или являются нечисловыми (например, текстовые данные). Также метод k-средних может быть менее эффективным в случаях, когда имеется большое количество выбросов или когда кластеры имеют неравную дисперсию.

Метод DBSCAN

Метод DBSCAN (Density-Based Spatial Clustering of Applications with Noise) – это алгоритм кластеризации, который находит плотные области в пространстве данных и определяет выбросы. Он может быть использован для кластеризации как числовых, так и категориальных данных [4].

Принцип работы метода DBSCAN заключается в поиске областей с высокой плотностью точек, которые являются кластерами. Для этого алгоритм использует два параметра: радиус ε и минимальное количество точек, необходимое для образования кластера (minPts).

Алгоритм начинается с выбора произвольной нерассмотренной точки в данных. Затем алгоритм находит все точки, которые находятся в радиусе ε от этой точки. Если количество точек в этой области больше или равно minPts, то эта область считается кластером. Если же количество точек меньше minPts, то эта точка считается выбросом. Если же количество точек больше minPts, но они не находятся в области с радиусом ε от выбранной точки, то алгоритм переходит к следующей нерассмотренной точке и повторяет процедуру.

После обнаружения первого кластера алгоритм находит все точки в данных, которые находятся в радиусе ε от этого кластера, и повторяет процедуру поиска кластеров для этих точек.

Одним из основных преимуществ метода DBSCAN является то, что он не требует заранее заданного числа кластеров и может определять их число на основе данных. Также метод DBSCAN может обрабатывать данные с шумом и выбросами, так как он может отличать их от настоящих кластеров.

Однако метод DBSCAN имеет некоторые недостатки. Например, если данные имеют разную плотность в разных областях, то параметры радиуса ε и minPts могут не подходить для всех кластеров. Также метод DBSCAN может иметь проблемы в случае, когда кластеры имеют сложную форму или пересекаются.

Метод DBSCAN широко используется в различных областях, таких как обработка и анализ данных, компьютерное зрение, биоинформатика и другие. Он может быть использован для кластеризации объектов, сегментации изображений, выявления аномалий и других задач.

Также данный метод применяется во многих областях, где требуется кластеризация данных, особенно в тех случаях, когда число кластеров заранее неизвестно. Вот некоторые задачи, где метод DBSCAN используется чаще всего:

Обработка изображений. Метод DBSCAN может использоваться для сегментации изображений, то есть разделения изображения на различные сегменты в зависимости от цвета, яркости или других характеристик. Например, для выделения объектов на изображении.

Биоинформатика. В биоинформатике метод DBSCAN может использоваться для кластеризации геномных данных, анализа белковых структур, выявления паттернов в биологических данных и многих других задач.

Маркетинг и реклама. Метод DBSCAN может использоваться для кластеризации покупателей на основе их покупательского поведения или для определения сегментов рынка.

Финансы. Метод DBSCAN может использоваться для кластеризации финансовых данных, таких как торговля на фондовых рынках, прогнозирование рисков и т.д.

Геоданные. В геоинформатике метод DBSCAN может использоваться для кластеризации пространственных данных, например, для выделения кластеров точек на карте.

Анализ социальных сетей. Метод DBSCAN может использоваться для кластеризации пользователей социальных сетей на основе их профилей или поведения.

Метод DBSCAN может использоваться во многих задачах, где требуется выявление структуры в данных и разделение их на кластеры.

Хотя метод DBSCAN имеет много преимуществ, таких как возможность определения количества кластеров и обнаружение шума, он также имеет некоторые недостатки, которые следует учитывать:

1. *Вычислительная сложность*. Метод DBSCAN имеет квадратичную сложность по отношению к количеству объектов данных. Это может стать проблемой для больших наборов данных или для задач, где требуется быстрый анализ в реальном времени.

2. *Чувствительность к параметрам*. Метод DBSCAN имеет два основных параметра – радиус epsilon и минимальное количество точек в кластере. Выбор правильных параметров может быть сложной задачей и зависеть от характеристик конкретного набора данных.

3. *Проблемы с выделением кластеров различных плотностей*. Метод DBSCAN не всегда хорошо работает с наборами данных, в которых кластеры имеют различные уровни плотности. В некоторых случаях он может выделить один большой кластер вместо нескольких меньших.

4. *Неэффективность на высокоразмерных данных*. Метод DBSCAN может столкнуться с проблемой «проклятия размерности», то есть снижения эффективности при работе с данными высокой размерности. Это может привести к тому, что метод DBSCAN не сможет найти кластеры в таких данных или даст неточный результат.

В целом метод DBSCAN является мощным инструментом для кластеризации данных, но его применение должно быть осознанным и учитывать особенности конкретной задачи и набора данных.

Методы кластеризации на основе плотности используются для идентификации кластеров в данных, основываясь на плотности распределения точек. Основная идея заключается в том, что кластеры представляют собой области с высокой плотностью точек, отделенные от других областей с низкой плотностью. В этом разделе мы рассмотрим два наиболее распространенных метода кластеризации на основе плотности: DBSCAN и OPTICS.

Метод кластеризации SOM и сеть Кохонена

Методы, основанные на многократном самоорганизующемся отображении (SOM), являются одним из методов кластеризации на основе нейронных сетей. SOM – это метод, в котором каждый элемент данных представляется точкой в многомерном пространстве и проецируется на двумерную сетку нейронов, которые формируют топологическую карту [5].

Процесс SOM начинается с инициализации случайных весов для каждого нейрона в сетке. Затем выбирается случайный элемент данных и определяется ближайший нейрон (winner neuron), т.е. нейрон, веса которого находятся ближе всего к данному элементу. Затем веса всех нейронов в окрестности победителя корректируются, чтобы они стали ближе к этому элементу данных. Этот процесс называется обучением.

Процесс обучения продолжается для каждого элемента данных в выборке. При этом размер окрестности нейронов, которые должны быть обновлены, постепенно уменьшается. В результате обучения происходит формирование топологической карты, где каждый нейрон соответствует определенному кластеру в данных.

Одним из основных преимуществ метода SOM является возможность визуализации кластерной структуры данных в двумерном пространстве, что позволяет более наглядно представить результаты кластеризации и проводить дальнейший анализ данных.

Методы, основанные на SOM, широко применяются в области обработки изображений и анализа данных, включая:

1. *Классификацию изображений*: SOM могут использоваться для классификации изображений по различным признакам, таким как цвет, текстура, форма и т.д.
2. *Рекомендательные системы*: SOM могут использоваться для кластеризации пользователей и товаров в рекомендательных системах, что позволяет предлагать пользователям наиболее подходящие товары.
3. *Анализ данных*: SOM могут использоваться для кластеризации больших наборов данных в различных областях, таких как финансы, медицина, биология и т.д.
4. *Обработка естественного языка*: SOM могут использоваться для кластеризации текстовых данных, таких как новости, статьи, отзывы и т.д.

Сеть Кохонена, также известная как карты Кохонена, или SOM (Self-Organizing Maps), является одним из методов кластеризации на основе нейронных сетей. Этот метод был разработан финским ученым Теуво Кохоненом в 1982 году и с тех пор получил широкое применение в различных областях, таких как обработка изображений, распознавание образов, биоинформатика и др.

Сеть Кохонена использует нейроны, расположенные в двумерной сетке, для представления многомерных данных. Каждый нейрон связан с определенным участком входного пространства, и каждый нейрон имеет свое значение веса, которое настраивается в процессе обучения. Обучение сети Кохонена происходит путем присвоения входных данных случайным весам нейронов и последующей корректировки этих весов в процессе обучения.

Основная идея метода заключается в том, чтобы сжать многомерное пространство входных данных в двумерную карту нейронов таким образом, чтобы близкие входные данные оказались в близких друг к другу нейронах на карте. Таким образом, каждый нейрон представляет определенный кластер данных.

Процесс кластеризации в сети Кохонена начинается с инициализации весов каждого нейрона случайными значениями, затем на каждом шаге происходит выбор случайного входного вектора и определение ближайшего нейрона на карте. Выбранный нейрон и его ближайшие соседи затем корректируют свои веса, чтобы стать ближе к выбранному входному вектору. Этот процесс повторяется многократно до тех пор, пока не будет достигнуто определенное количество эпох обучения.

Сеть Кохонена часто используется для визуализации данных в двумерном пространстве, где каждый нейрон представляет определенный кластер данных. Это позволяет легче интерпретировать результаты кластеризации и понимать структуру данных. Также сеть Кохонена может использоваться для классификации и прогнозирования, особенно в задачах, связанных с обработкой изображений и сигналов.

Кластеризация с помощью нейронных сетей

Методы кластеризации, использующие нейронные сети, представляют собой разновидность методов машинного обучения, которые основываются на алгоритмах нейронных сетей. Нейронные сети представляют собой математические модели, которые имитируют работу мозга и способны обрабатывать сложные данные. Кластеризация с помощью нейронных сетей обычно выполняется в два этапа: обучение и классификация [6].

Обучение нейронной сети заключается в том, чтобы настроить ее параметры, чтобы она могла определить структуру данных и выделить кластеры. Это можно сделать, например, путем обучения с учителем, где данные размечены заранее известными метками кластеров, или с помощью обучения без учителя, где нейронная сеть сама ищет закономерности в данных.

Классификация с помощью нейронных сетей заключается в том, чтобы использовать обученную нейронную сеть для определения принадлежности точек кластерам. Для этого данные передаются через обученную нейронную сеть, которая выдает вероятности принадлежности точек кластерам. Затем можно использовать пороговое значение вероятности для определения, к какому кластеру относится каждая точка.

Существует несколько методов кластеризации, использующих нейронные сети, включая методы, основанные на многократном самоорганизующемся отображении (SOM), нейронных сетях Кохонена, глубоком обучении и других. Каждый метод имеет свои преимущества и недостатки и может быть эффективным в различных задачах кластеризации.

Одним из преимуществ методов кластеризации, использующих нейронные сети, является их способность обнаруживать сложные зависимости в данных и работать с высокоразмерными данными. Кроме того, нейронные сети могут обучаться на больших объемах данных и обрабатывать информацию в режиме реального времени [7].

Однако методы кластеризации, использующие нейронные сети, могут быть вычислительно сложными и требовать большого объема вычислительных ресурсов. Кроме того, обучение нейронных сетей может требовать большого количества данных и времени на обучение. Также при использовании нейронных сетей для кластеризации могут возникать проблемы с интерпретируемостью результатов, что затрудняет понимание структуры данных и интерпретацию полученных результатов.

Стоит отметить, что выбор метода кластеризации, включая методы на основе нейронных сетей, зависит от специфики данных и задачи, которую необходимо решить. В некоторых случаях лучше могут работать методы на основе плотности, в других – методы на основе расстояний, а в третьих – методы на основе нейронных сетей.

Несмотря на ограничения и проблемы, методы кластеризации, использующие нейронные сети, являются мощным инструментом для обработки и анализа данных. Они могут помочь выделить скрытые закономерности и структуры в данных, что может быть полезным в различных областях, включая медицину, финансы, науку о материалах и другие. Кроме того, с развитием технологий и возможностей вычислительной техники методы кластеризации, использующие нейронные сети, становятся все более доступными и применимыми в реальных приложениях.

Методы кластеризации на основе нейронных сетей широко применяются во многих областях:

1. *Обработка изображений:* нейронные сети могут использоваться для выделения объектов на изображениях и их классификации по кластерам в зависимости от различных признаков, таких как цвет, форма, текстура и т.д.

2. *Обработка звука:* нейронные сети могут использоваться для классификации и кластеризации звуковых сигналов, таких как речь, музыка, звуки окружающей среды и т.д.

3. *Анализ данных:* нейронные сети могут использоваться для кластеризации больших наборов данных в различных областях, таких как финансы, медицина, биология и т.д.

4. *Обработка естественного языка:* нейронные сети могут использоваться для кластеризации текстовых данных, таких как новости, статьи, отзывы и т.д.

5. *Рекомендательные системы:* нейронные сети могут использоваться для кластеризации пользователей и товаров в рекомендательных системах, что позволяет предлагать пользователям наиболее подходящие товары.

6. *Биоинформатика:* нейронные сети могут использоваться для кластеризации биологических данных, таких как геномы, протеомы и т.д.

Это лишь несколько примеров того, где и как могут быть использованы методы кластеризации на основе нейронных сетей. В целом они могут быть полезны в любой области, где необходимо выявить скрытые закономерности и структуры в данных.

Использование нейронных сетей для кластеризации данных имеет свои преимущества и недостатки.

Одним из главных преимуществ является возможность обработки больших объемов данных с высокой точностью. Нейронные сети могут обрабатывать сложные данные, которые не могут быть решены с помощью традиционных методов кластеризации. Кроме того, нейронные сети могут выделять скрытые зависимости между данными, что поможет в создании более качественных кластеров.

Однако недостатки тоже есть. Например, настройка параметров нейронной сети может быть сложной и требует опыта. Также нейронные сети часто требуют большого количества данных для обучения, что может быть проблематично в некоторых приложениях.

Заключение

В целом использование нейронных сетей для кластеризации данных является эффективным подходом, который может помочь в создании качественных кластеров, особенно в случаях, когда объемы данных большие и данные сложны для анализа традиционными методами кластеризации. Однако выбор конкретного метода кластеризации должен зависеть от конкретной задачи и требований к точности и скорости работы.

СПИСОК ЛИТЕРАТУРЫ

1. *Осовский С.* Нейронные сети для обработки информации. Москва: Финансы и статистика, 2016.

2. *Мандель И. Д.* Кластерный анализ. Москва: Финансы и статистика, 1988. 176 с.

3. *Raghavan R.* A fast and scalable hardware architecture for K-means clustering for big data analysis : University of Colorado Colorado Springs. Kraemer Family Library, 2016.

4. *Kriegel H.-P., Schubert E., Zimek A.* The (black) art of runtime evaluation: Are we comparing algorithms or implementations? Knowledge and Information Systems. 2016. Vol. 52. No. 2. P. 341.

5. *Kohonen T.* Self-Organizing Maps (Third Extended Edition). New York, 2001. 501 p.

6. *Вятченин Д. А.* Нечеткие методы автоматической классификации. Минск: Технопринт, 2004. 219 с.

7. *Жилов Р. А.* Применение нейронных сетей при кластеризации данных // Известия Кабардино-Балкарского научного центра РАН. 2021. № 1(99). С. 15–19.

REFERENCES

1. Osovsky S. *Neyronnyye seti dlya obrabotki informatsii* [Neural networks for information processing]. Moscow: Finansy i statistika, 2016. (In Russian)

2. Mandel I.D. *Klasternyy analiz* [Cluster analysis]. Moscow: Finansy i statistika, 1988. 176 p. (In Russian)

3. Raghavan R. A fast and scalable hardware architecture for K-means clustering for big data analysis : *University of Colorado Colorado Springs.* Kraemer Family Library, 2016.

4. Kriegel H.-P., Schubert E., Zimek A. The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowledge and Information Systems*. 2016. Vol. 52. No. 2. P. 341.

5. Kohonen T. Self-Organizing Maps (Third Extended Edition). New York, 2001. 501 p.

6. Vyatchenin D.A. *Nechotkiye metody avtomaticheskoy klassifikatsii* [Fuzzy methods of automatic classification]. Minsk: Technoprint, 2004. 219 p. (In Russian)

7. Zhilov R.A. Application of neural networks in data clustering. *News of the Kabardino-Balkarian Scientific Center of RAS.* 2021. No. 1(99). Pp. 15–19. (In Russian)

Информация об авторе

Жилов Руслан Альбердович, мл. науч. сотр. отдела нейроинформатики и машинного обучения, Институт прикладной математики и автоматизации – филиал Кабардино-Балкарского научного центра Российской академии наук;

360000, Россия, г. Нальчик, ул. Шортанова, 89 А;

zhilov91@gmail.com, ORCID: https://orcid.org/0000-0002-3552-4854

Information about the author

Zhilov Ruslan Alberdovich, Junior Researcher of the Department of Neuroinformatics and Machine Learning, Institute of Applied Mathematics and Automation – branch of Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences;

360000, Russia, Nalchik, 89 A Shortanov street;

zhilov91@gmail.com, ORCID: https://orcid.org/0000-0002-3552-4854